DOCUMENT RESUME

ED 305 391 TM 012 942

AUTHOR Daiker, Donald A.; Grogan, Nedra
TITLE The Selection and Use of Sample Papers in Holistic Evaluation.
PUB DATE 85
NOTE 33p.
PUB TYPE Reports - Research/Technical (143) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Advanced Placement; *Essay Tests; Evaluators; Higher Education; *Holistic Evaluation; *Interrater Reliability; Scoring; Test Construction; Undergraduate Students; *Writing Evaluation
IDENTIFIERS Educational Testing Service; Placement Tests; *Sample Materials; *Writing Samples

ABSTRACT

The role of sample papers (i.e., anchor papers, prototypes, range-finders) in holistic evaluation of writing is discussed. When, where, and how many sample papers are to be selected, and who should perform the selection are covered. The process of sample selection should proceed as follows: (1) a general reading of papers by committee members to provide preliminary scoring of papers; (2) proposal and review of sample papers by committee members; (3) establishment of papers on which consensus scores have been reached; (4) review of papers for which consensus is low; and (5) selection of two or more sample papers for each point on the scoring scale. Procedures described are generally those used by the Educational Testing Service for readers of advanced placement examinations. The central principle in sequencing papers is to guide readers into making increasingly more difficult and sensitive scoring decisions. Finally, a statement of scoring criteria should be developed; this rubric describes and characterizes papers at each end of the scoring scale. A description of a recent holistic evaluation session at Miami University (Oxford, Ohio) illustrates the use of sample papers in training of raters and scaling of papers. Sample papers are included in the discussion. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED305391

The Selection and Use of Sample Papers in Holistic Evaluation

Donald A. Daiker and Nedra Grogan

Miami University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DONALD A. DAIKER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Donald A. Daiker
Department of English
Miami University
Oxford, OH 45056

(513) 529-7110 (O)
(513) 523-2732 (H)

Nedra Grogan
Department of English
Miami University
Oxford, OH 45056

(513) 529-5221 (O)
(513) 523-6084 (H)

TM012942

The Selection and Use of Sample Papers

in Holistic Evaluation

During a holistic scoring session several years ago, a table leader returned from collecting sample paper scores, leaned forward in her chair, and said softly to the seven readers at her table, "We've got real trouble here."

The trouble, the table leader explained, lay in the readers' scoring of a sample paper. Using the standard 9-point scale, three readers had scored the sample a "6," three others had scored it a "5," and one reader had given it a "4"--not an unusual or especially troublesome distribution in itself. The problem, the table leader announced, was that the paper was offered to illustrate the score of "8." Her announcement was greeted with incredulity: "An '8'?" "You've got to be kidding!" "There's no way that paper deserves an '8'!" Intense discussion followed, and it continued long after other tables had resumed scoring "live" papers. Finally, the question leader felt it necessary to intervene. "Look," he said in his most teacherly tone, "if this table can't wind up its discussion, perhaps it had better continue out in the hall." His message was to be quiet and get back to work. But one reader, not catching on, responded, "Yeah, let's get this thing settled outside." So the table leader and her seven readers paraded past the inquisitive eyes of dozens of other readers in order to

continue discussing a sample paper out in the hall. But that didn't help much, either. The readers returned to scoring student essays genuinely confused, at least until the reassurance of the next sample paper, about the scoring criteria they were expected to employ.

Of course, it doesn't take a reader walkout to convince anyone experienced in holistic scoring of the vital importance of sample papers. By general consensus they are, next to the chief reader, the single most important element in a successful scoring session. But if a test developer or chief reader looks to scholarship for help in selecting and using sample papers, she will not find much. The literature on holisticism, for example, is strangely similar to the actual procedure: both the method and what is written about it are "holistic," with very little attention given to individual parts. A crucial missing part in the "whole" of holisticism is the selection and use of sample papers. In the literature on writing assessment in general and holistic evaluation specifically, the selection and use of sample papers is a lost issue among larger concerns of reliability and validity or among more "obvious" concerns like writing prompts.

Perhaps the variety of names for sample papers--anchor papers, prototypes, range-finders--suggests their multiple characteristics and manifold functions. A review of the literature shows that sample papers are supposed to fit all of the following descriptions: papers with attributes that

characterize each score level and exemplify various point scores (Cooper, 1984, p.41); papers which have "the desired normative distribution" (Myers, 1980, p. 32); papers that are representative, especially of the middle range of a single score (Northwest, 1980, p. 65); papers "which illustrate the full range of performance on the assignment" (Odell & Cooper, 1980, p. 36); papers that "virtually all readers" will agree upon (Spandel & Stiggins, 1980, p. 20); and papers that illustrate "the distinctions in quality described by the scoring guide" (White, 1985, p. 160). While these are overlapping descriptions, the differences among them complicate the role of sample papers while indicating very little about selecting and using them.

Those who write about assessment are able to sidestep the troubling nature of sample papers because explanations of procedure usually must be carefully qualified: "It just depends on the particular testing situation." Despite the appropriateness of such a guideline--and its usefulness will become evident as we continue--it cannot serve as the answer to all the critical questions which precede a reading. Our purpose here is to outline the series of decisions involved in selecting and then using sample papers. We want, in particular, to make test-developers aware that they have choices to make and that their choices will have serious consequences for student and evaluator alike.

SELECTING SAMPLE PAPERS

Although sample papers have come to seem an integral part of holistic evaluation, the two did not develop simultaneously. Holistic evaluation came first, followed by sample papers only after researchers found readers to be unreliable without them. Coffman and Kurfman (1968) were among the first to recognize the "idiosyncratic fluctuations" of readers in both analytic and holistic scoring situations. They conclude--and their recommendation has been embraced by the Educational Testing Service and widely accepted by the profession--that sample papers must be used, regularly and repeatedly, to train readers in holistic evaluation.

The procedure of selecting sample papers to train readers begins with relatively uncomplicated matters of logistics and planning. When sample papers are selected depends, first, on when the test results are needed and, second, on when the selection committee can be brought together. Unless the writing assignment has been fieldtested earlier (always a sound practice), the session for selecting papers cannot take place until the writing has been completed. Because results are often wanted immediately, especially for placement and proficiency examinations, the selection session is commonly held as soon after the writing as the selection committee members can assemble. Where should the selection session take place? Any quiet setting with a table large enough to spread out stacks of papers and with chairs comfortable

enough for colleagues who may occasionally be made uncomfortable by the judgments of others.

Who should select sample papers? The most important principle is that the selection process requires more than a single person. A chief reader invites trouble--and perhaps even disaster--by selecting sample papers on her own. At some point during the subsequent training session, she can be sure that her evaluation of a sample paper will be challenged and that she will have to explain and defend her score. Her explanation will be more persuasive if readers are reminded that the score represents the consensus of the full selection committee and if committee members follow with supporting comments of their own. In our experience, a five-member selection committee is ideal, three is the bare minimum, and more than seven becomes unwieldy. The second important principal concerns the choice of committee members, and here common sense prevails: taking into account local political realities, look for members who have experience in holistic evaluation, who have earned the trust and professional respect of their colleagues, and who do not define compromise as loss of face.

The ultimate goal of sample paper selection is reader consistency. The immediate goal is the identification of a set of papers to be used in training readers to reach consistency. How many papers are needed depends to some extent on paper length--shorter papers take up less training time--but more importantly on the range of the scoring scale

and the number of reading days. You will need more papers to illustrate the 9-point scoring scale used by the Educational Testing Service to score Advanced Placement essays than for the 4-point scale used by several universities for placement purposes. By the same token, ETS requires many more samples for its six-day reading than Miami University needs for a reading that takes only half a day. With the popular 6-point scale, two papers for each point (a total of 12) are usually more than enough for a single day of reading. For each additional day, select another paper for each point of the scale.

How do you get 12 or 18 sample papers? The first step is for the selection committee members to read papers from the batch to be evaluated. If papers have been randomized, each member grabs 40 to 50 papers and retreats to a quiet spot. If the papers have not been randomized, and especially if they are arranged by class, school, or ability level, the chief reader provides each committee member with a roughly representative set of papers--papers that can be expected to approximate the range of the entire population.

With a set of papers in hand, committee members start reading on their own. And that's all they do--read. They temporarily forget about scores and scoring scales. That's because the primary purpose of this initial reading is to get a general sense of how students have responded to the assignment: to begin observing differences between more and less successful papers, and to begin developing a feel for

the range of responses. It is important that these first papers not be evaluated in terms of a scoring scale, that they not be assigned a score or placed in piles corresponding to the points of the scoring scale. Since holistic scoring is nothing more than the rank ordering of papers from one particular testing situation, scoring can not begin--and should not be attempted--until a number of papers from that testing situation have been read.

Only after an hour or more of reading--without scoring--should each committee member move, on her own, to the second step in the selection process: the preliminary scoring of papers. Put aside the papers you've already read--they've served their purpose--and take new ones from your set. Now is the time to begin applying the scoring scale--deliberately and thoughtfully. But do not write on the papers. Instead, place each into a pile according to the score you've assigned it; for a six point scale there will be six piles. When you have two or three papers in each pile--you sometimes have to hunt for papers at the ends of the scale--you are ready to review your scoring. Start at either the upper or lower end of the scale and reread, in order, the papers in each pile, making any necessary adjustments as you go. Your aim is to choose, for each point of the scale, one or two papers to recommend as sample papers to the selection committee.

When the selection committee meets--and this is step 3--each member brings the papers he is proposing as samples. The chief reader begins the formal part of the meeting by

asking everyone to place a single, unmarked paper in the middle of the table. The chief reader makes clear--and this is a crucial point--that each member is free to choose a paper from any point on the scoring scale. The more points represented, the better. Problems lie in the other direction--when the chief reader invites members to submit, say, samples of upper-range papers. Such an invitation may speed up the selection process but it can only bias the results. As Paul Diederich (1974, p. 12) reminds us, "grading is such a suggestible process that we find what we expect to find." If we are told to expect upper-range papers, you can bet we'll find them. The best way to limit bias, of course, is not to create expectations.

From the pile of proposed sample papers, each member takes one, reads and scores it, silently records the score and paper number on a separate slip of paper, and then repeats the process until all the papers are scored. Scoring slips are now passed to a designated committee member, who reads the scores aloud so that the chief reader can record them on a chalkboard for all to see and consider. Once the scores for all papers have been tabulated, discussion begins.

It's a good idea for the chief reader to begin discussion by focusing on the paper that comes closest to gaining a consensus score. If one of the proposed samples receives, say, a "3" from everyone, that's the paper to start with. The chief reader announces the scores and then asks a question like, "Why did you score this a '3'?" The responses

will be useful in several ways: they provide the chief reader (who is taking notes) with material for the subsequent training session; they begin the process leading to a written statement of criteria (sometimes called a rubric); and they make it easier to select sample papers representing other points on the scoring scale. Unless the responses unexpectedly reveal confusion or disagreement, the paper that everyone graded a "3" is then set aside as one of the samples to be used during the training session.

If other papers have received the same score from all members, they can be discussed next. It makes good sense to move, in order, from the papers that have produced the most initial agreement to papers that have produced the least. Of course, a paper need not have received exactly the same initial score from all members to qualify eventually as a sample paper. But a cautious chief reader will not select as a sample paper one that has received more than a single divergent score or one whose single divergent score is off by more than a point. Nor will she choose a paper with a single, one-point divergent score unless the divergent scorer willingly changes his mind as a result of discussion.

From the first-half dozen papers proposed as samples, at least one usually gains consensus and is accepted. But the proposed samples that fail to gain consensus are absolutely essential to the selection process. These papers need to be discussed fully--a good five minutes of discussion for each--in order to lay the basis for future agreement. The lack of

consensus is a sure sign that committee members are not yet employing similar standards and criteria; discussion of specific papers is the most tactful and effective way of making us aware of the situation and of encouraging us to come together. Not that we are always willing to change our evaluation of a paper: it is not easy to admit that we were "off" or that we "missed one." But when we recognize that our evaluation varies significantly from that of colleagues we respect, we silently make the kinds of adjustments that make future agreement possible.

The process of submitting, reading, scoring, and discussing papers continues until two or more sample papers have been selected--by consensus--for each point of the scoring scale. Once the papers have been selected, it's time for the committee's final three decisions: how many sample papers at a time should be given to readers during the training session? In what sequence should they be presented? And finally, how should leaders describe in writing the scoring criteria?

How many at a time depends in large part on paper length. ETS training sessions for readers of AP examinations in English typically begin with four papers--a "2," a "4," a "6," and an "8" on the nine-point scale. But AP essays are written in 35 or 40 minutes and tend to be short, averaging just over 200 words. Even so, subsequent batches of papers at AP training sessions usually consist of two or three papers, never more than four. With longer papers, especially

those in the range of 500 words, it seems more reasonable to use pairs of papers from the beginning.

The central principle in sequencing papers is to guide readers into making increasingly more difficult and sensitive scoring decisions. For this reason, it makes sense to start with a pair of papers that present a clear contrast in quality without either paper representing the very top or the very bottom of the scoring scale. With the 6-point scale, we like to begin with a "5" and a "2," or a "5" and a "3." Either pair, but especially the "5" and "2," is likely to produce consensus--at least in terms of which is the better of the two papers. It also makes sense to begin with papers that represent the middle of both the top range (4-6) and bottom range (1-3). Once readers experience mid-range papers, they can more readily identify papers at the ends of the scale.

If the initial pair is a "5" and "2," the committee might choose a "6" and a "3" or a "4" and a "1" as its second pair. In either case, it makes sense to choose papers that have earned a "solid" score--until each point on the scoring scale has been represented once. A solid "5" or a solid "2" is one that committee members agree occupies the middle of the "5" or "2" range (Northwest, 1980, p. 65). (As this statement implies, each point on any scoring scale represents a range of scores; that's why teachers are fond of adding a plus or minus to the letter grades they assign.) Once "solid" papers have been chosen to illustrate every point on

the scoring scale, the committee next decides whether to include papers judged to occupy the bottom or top of its point range (a high "3," a low "5"). A final decision is whether to use borderline papers (a 3/4 split; a 5/6 split) in the training session that follows.

After sample papers have been placed in order of use and turned over to be labeled, photocopied, and bound in a packet for the upcoming training session, leaders are left with a final decision: they must put together a rubric, a statement of scoring criteria. The rubric describes and characterizes papers at each end of the scoring scale (See pg. 17 for a sample rubric). Creating a rubric must be the very last step of sample paper selection, for until papers have been chosen and discussion completed, no scoring criteria exist.

Without the advantage of the discussion that accompanies the selection of sample papers, a rubric will be imposed rather than "inferred," and most researchers agree that scoring criteria must be allowed to develop inductively from a particular set of papers (Myers, 1980; Odell and Cooper, 1980). Because of the very nature of holistic scoring--a procedure limited to only a rank ordering of papers from a single testing situation--the scoring criteria must evolve "naturally" from the testing situation and from committee members' dialogue about each score level. Experienced readers' choices of representative writing samples become the foundation of the scoring guide.

Sometimes, of course, a chief reader is not able to make

use of the full selection process described here. Sometimes constraints of time and money force her to make procedural compromises. The least damaging compromise, in our view, is simply to eliminate the first two steps above: the chief reader, and not the committee members, reads widely among the essays and makes a preliminary selection of sample papers. These papers are then blindly scored by the committee members and the chief reader's preliminary scoring is either accepted or rejected, paper by paper. The most damaging compromise is for the chief reader to select samples on her own or to bias the reading of the selection committee with directions like "See if you don't agree that this is a solid "4" paper." Compromises are more difficult to make at the next stage of the holistic procedure--the use of sample papers.

USING SAMPLE PAPERS

Despite its complexities, the selection of sample papers is only preliminary to their use in a training session. Since the major goal of training is to guide readers toward consistency in evaluation, the chief reader tries to use sample papers in such a way as to achieve reader agreement--or calibration--quickly and harmoniously.

Even before the first sample paper is distributed, the chief reader and committee members will have opportunities

for creating an environment conducive to consensus. The choice of a well-lighted and comfortable room, the availability of snacks and beverages, a warm welcome and personal introductions, and perhaps even reader involvement in preliminary tasks of sorting and distributing--all these contribute to the sense of collegiality that makes agreement possible. Just as important, because it builds confidence in both the leaders and the procedures, is a sense of professionalism that can be established through seemingly insignificant details like name cards at assigned seats, designated reader numbers, coded papers, and a printed schedule of the day's activities.

The choice of a room for the reading requires more planning than a call to the room-reservations office. Leaders will want to decide in advance between tables and theatre seating. If tables are used, as ETS does, committee members commonly become table leaders. They assume primary reponsibility for using sample papers to train and calibrate the five-to-ten readers seated at their table. But for small holistic sessions, theatre seating is more efficient, more economical, and perhaps equally effective. With theatre seating, the chief reader stands at the front of the room with the readers seated at desks in a semi-circle before her.

Once the setting has been established and a collegial atmosphere created, the formal training session begins. To illustrate the kinds of decisions the training session

presents for a chief reader--especially decisions involving sample papers--we will describe a recent holistic evaluation session at Miami University. Here is the writing assignment:

COLLEGE COMPOSITION PROFICIENCY EXAMINATION
FALL 1988

Instructions:
Spend about 20-30 minutes reading and thinking about the poem "Next Day" by Randall Jarrell. During this time, you may want to free write, take notes, construct an outline or write a partial draft. When you feel you've come to an understanding of the poem, write an essay of at least 500 words explaining how the poetry works to create your impression of the speaker. You may wish to evaluate this character in light of your own experiences or in relation to characters you know from movies, tv, or other works of literature. You may also consider how this character reflects the poet's values or attitudes. Whatever your views of the character, they should grow out of a careful examination of the text, including features such as voice, figurative language, and sentence structure.

Criteria for Evaluation:
The instructors of composition and literature who evaluate your essay will be looking for evidence of your ability to read intelligently and insightfully and to write clearly and persuasively. In particular, they will be looking for:

1. A unified, well-organized analysis offering a persuasive interpretation of the poem;
2. An essay rich in detail--one whose key assertions are supported by specific references to the text; and,
3. A well-written essay, demonstrating effective word choice, sentence structure, paragraph construction, and control of other writing strategies.

A Two-Hour Examination:
The exam lasts a full two hours. So as not to disturb others, you will not be permitted to leave the examination room until those two hours have elapsed. Since you must be here anyway, it makes good sense to use the full two hours to think about, plan, write, rewrite, revise, and proofread your essay.

NEXT DAY
by Randall Jarrell

Moving from Cheer to Joy, from Joy to All,
I take a box

And add it to my wild rice, my Cornish game hens.
The slacked or shorted, basketed, identical
Food gathering flocks
Are selves I overlook. Wisdom, said William James,

Is learning what to overlook. And I am wise
If that is wisdom.
Yet, somehow, as I buy All from these shelves
And the boy takes it to my station wagon,
What I've become
Troubles me even if I shut my eyes.

When I was young and miserable and pretty
and poor, I'd wish
What all girls wish: to have a husband,
A house and children. Now that I'm old, my wish
Is womanish:
That the boy putting groceries in my car

See me. It bewilders me he doesn't see me.
For so many years
I was good enough to eat: the world looked at me
and its mouth watered. How often have they undressed me
The eyes of strangers!
And, holding their flesh within my flesh, their vile

Imaginings within my imagining,
I too have taken
The chance of life. Now the boy pats my dog
And we start home. Now I am good.
The last mistaken,
Ecstatic, accidental bliss, the blind

Happiness that, bursting, leaves upon the palm
Some soap and water--
It was so long ago, back in some Gay
Twenties, Nineties, I don't know. . . . Today I miss
My lovely daughter
Away at school, my sons away at school,

My husband away at work--I wish for them.
The dog, the maid,
and I go through the sure unvarying days
At home in them. As I look at my life,
I am afraid
Only that it will change, as I am changing:

I am afraid, this morning, of my face.
It looks at me
From the rear-view mirror, with the eyes I hate,
The smile I hate. Its plain, lined look
Of gray discovery
Repeats to me: "You're old." That's all, I'm old.

And yet I'm afraid, as I was at the funeral
I went to yesterday.
My friend's cold made-up face, granite among its flowers
Her undress, operated-on, dressed body
Were my face and body.
As I think of her I hear her telling me.

How young I seem; I am exceptional;
I think of all I have.
But really no one is exceptional,
No one has anything, I'm anybody,
I stand beside my grave
Confused with my life, that is commonplace and solitary.

When the writing assignment involves analysis and interpretation of a piece of extended writing--an essay, a short story, a long poem--it's sound practice to provide readers with the text at least a day in advance of the reading. But even when readers receive the text beforehand, it makes sense to allow several minutes for rereading and reviewing. The chief reader then invites questions about the instructions and--if she chooses--about the text itself. Although it is important for the chief reader to make sure readers understand what the writing assignment asks of students, a discussion of the poem's meaning at this point will tend to shift the focus to interpretation--and away from evaluation. So it may be wiser to encourage only questions that relate to the instructions or the testing situation.

The chief reader's last task before turning to sample papers is to briefly define the rubric--the scoring scale and rating criteria. All readers should receive their own copy of the rubric so that they can refer to its descriptions

both during the training session and when scoring live papers. Here is a statement of the rating criteria Miami's chief reader and committee members designed after selecting sample papers but before the training session:

Rating Criteria
College Composition Proficiency Exam - Fall, 1988

These are features characteristic of papers at each level:

6 — The best papers take charge of the assignment, offering a range of approaches to the poem. They are generally longer, including numerous interpretive insights and often showing the writer's ability to incorporate materials from outside the poem into a coherent essay. They are stylistically felicitous, offering satisfying turns of phrase or interesting sentence structures.

5 -- These writers likewise offer interesting and insightful interpretations of the poem, but may be less sophisticated in their handling of organization or theme. Like the 6 paper, this level includes abundant supporting references to the poem without digression or extended summary. The style is smooth, sometimes artful.

4 -- This rank represents the better performance in the middle range, but these papers fall below the best because of their unevenness. The writer may be less sensitive to the character; personal experience or reference to other texts may not work well with the interpretive paragraphs. Or the paper may lack balance: the reading may be impressive but the organization too predictable. But these writers will still offer a sufficient number of references to the text. While these papers are generally mechanically correct, the style can be less than elegant.

3 -- At this level, the papers are clearly unsuccessful--as interpretations and as compositions. The interpretation of the poem will often read like a summary with few specific references to the text. Use of conventional literary critical tools may seem labored. Organization may be flawed or predictable. Style may be overly simple or marred by odd diction.

2/1 -- At the lowest level, students usually have much less to say. Content is marked by repetition of empty phrases. The writer may rely heavily on summary. Organization may be tediously formulaic or random. While few of the papers will contain glaring or numerous mechanical errors, most will show

a lack of control of sentence structures and a limited vocabulary.

The rating criteria are read aloud, slowly, by the chief reader, who then turns to the heart of the training session: the distribution and scoring of sample papers. For each sample, readers are asked to read quickly and holistically--that is, by attending to the effect of the whole rather than to any individual part. Once they have arrived at a score, they are required to write it down; that way, they will be less likely to report a different score when the chief reader collects votes. Whether scores are collected by a show of hands or by slips of paper, they should be recorded on a chalkboard or overhead projector so that readers can sense the level of agreement (or disagreement) on each sample paper. When readers see, for example, that all their scores landed in the lower range, they also see evidence of shared judgments and common goals. From this "success" in judging a paper, the reading takes on an important atmosphere of community.

Finally, after all the logistical decisions have been made, the chief reader can distribute the samples and start the training session. Ideally, the samples have been photocopied, alphabetically labeled in tentative order of use, and stapled together as a single packet, one for each reader. Handing out packets rather than individual papers saves distribution time and keeps aides from distracting trips around the room. The chief reader can begin, then, by

asking the group to read and score papers "A" and "B" in their packet.

The packet of Miami University sample papers began with a "5" (paired with a "2). The following paper illustrates the advantages of starting with a sample in the upper-range:

SAMPLE PAPER "A"

"Donna Reed's Mid-life Crisis"

In the poem "Next Day" Randall Jarrell vividly expresses the thoughts and anxieties of a middle-aged, middle income woman. With flittering remembrances of dreams long dead and constant fears of death creeping forward, Jarrell paints an honest and insightful picture of growing old. Through the poet's use of voice, sentence structure, and language, she creates the true character.

In the first two stanzas, the reader is abruptly plunged into the ritualism of domesticity: grocery shopping. In the first three lines, "Moving from Cheer to Joy to All,/I take a box/And add it to my wild rice, my Cornish game hens.", the character appears as the ultimate Donna Reed, continuing the monotony of daily life. She thinks herself different, however, as she attempts to exclude herself from the blandness and the ordinary person; but she cannot. "Yet Somehow, as I buy All from these shelves / And the boy takes it to my station wagon, / What I've become / Troubles me even if I shut my eyes." She has been absorbed into the rat race of life and can't escape.

In the third and fourth stanzas, the character drifts into the dreams of the past. When she was young and poor, she wanted what she has achieved; a home, family, and husband: the blocks on which society is built. Now that she has attained these dreams, she wants her treasures of the past; attention, admiration, and love: the baseness of wildness and burning lust. For most of her life she has had "The eyes of strangers" ogling her, undressing her. Now, the bag boy won't even notice her. She wants the desire and sex, to be a seductress, but is deprived.

As she leaves for home, in the fifth stanza, the bag boy petting the dog somehow breaks the lustful spell; she returns to a pillar of society. "Now I am good." Her last sexual fantasy leaves only some soap and water on her hands--the staples of domesticity. Her youth is now something long forgotten. She longs for her children at school, her husband at work, rather than her days of jeune fille and freedom. She moves through her peaceful days with fear; she fears loss of security, of change in her great plane of calm. She

realizes that she is changing and fears so will her world.

As the eighth stanza begins, she is in disgust of her appearance--a complaint many women have. "... <T>he eyes I hate, / The smile I hate. Its plain lined look / Of gray discovery / Repeats to me: 'You're old.' That's all, I'm old." The ninth stanza gives some basis of what brought this mid-life crisis--the death of a friend. As she was at the funeral, she saw herself as the "...cold made-up face, ...". She now has the realization that she is mortal and she is getting older. She thinks of her friend speaking to her, telling how young she looks and how exceptional she seems. But, she feels she is not exceptional, she is just another face in the crowd, a part of the masses. She is closing in on death and feels her life is commonplace and solitary. She is alone.

In "Next Day" (the day after the funeral), the main character has undergone tremendous stress and fears. She fears aging, death, and being alone. The voice of the character is one of flighting fantasy that must return to the morals of a pillar of society. It's like Donna Reed having an intense sexual dream; it doesn't last long. The language shows anxiety and monotony coexisting. While she is bored with her life, she doesn't want it to change it is her security against the world. She is confused, tired, and afraid. She is middle-aged.

The chief reader's decision to begin with an upper-range paper, particularly this one, was confirmed when it elicited high agreement from the readers: 46 raters out of 64 scored it a "5." The paper worked well as a first sample mainly because of its memorable quality; most readers saw the Donna Reed analogy as its best feature and applauded the paper's skillful language as well as its "fun moments." For readers to evaluate this sample paper first--and to see high agreement on the score of "5" recorded on the chalkboard--helped them define the upper range and, as a result, enabled them more accurately to identify papers in the middle and lower ranges. When readers were next given the paired paper selected as a "2," fully half of them scored it a "2," and

all agreed that it was inferior to the "5" sample.

Before the chief reader moves to a second pair of sample papers, she takes into account the level of agreement on the first. With acceptable agreement on the first pair--almost all scores within a single point of each other--she can feel comfortable in moving to sample papers representing points other than "5" or "2" on the scoring scale. But if the first pair of papers receives a wide range of scores, the chief reader may want to choose a second pair of papers that the leaders had scored "5" and "2."

The choice of which pair to use next may depend on the context of the testing situation. In the case of the Miami University proficiency examination, it was clear to the leaders that the distinction between a "4" and a "5" would be a crucial one. Students who earned the grade of "5" from two readers would most likely earn exemption from college composition as well as six hours of course credit; but students who twice received the score of "4" would probably be required to enroll in one semester of college composition and receive no course credit. Given the importance of defining papers in the top range, Miami's chief reader chose for the second sample set a pair of papers that leaders had judged to be a "6" and a "4." Here is the sample paper that leaders had scored a high "6":

SAMPLE PAPER "C"

"Next Day" by Randall Jarrell is a poignant account of an ordinary woman's discovery of the inevitability of growing

old. This work begins with Jarrell's character in a grocery store where she is choosing soap. I believe the author uses soap as a thread throughout the poem which explains the death of youthful invincibility. This woman has moved from the cheer and joy of youth and chosen to take all in life as in the detergent aisle. She bought into all--the husband, the house, the children--she desired when she was young. The character has taken what Jarrell calls the chance of life. She fell in love and got married and the soap analogy comes to its fruition in line thirty. "Ecstatic, accidental bliss, the blind happiness that, bursting, leaves upon the palm some soap and water --". What is a bubble--that magical, glistening, rainbow of magic--but a bit of soap and water when it bursts.

My impressions of the speaker are mixed since I feel she has learned to overlook some aspects of her current situation while succumbing to other aspects. She is buying food which relates directly to domesticity, cooking, and family responsibility. In fact, all the women in the grocery are refered to as food gathering flocks who are basketed and identical in lines four and five. The speaker doesn't seem to be too concerned with the fact she has traded her individuality and lack of domesticity for all. She claims enough wisdom to overlook these aspects of her lot in life. Flirtation and admiration from others is a trade, however, that she still regrets. Her reference to the grocery carrier as a boy points out her increased age while the station wagon she drives is an ultra practical, fuddy duddy symbol of the sacrafice of admiring glances for family and children.

She is most troubled by her increasing age and the impact this has on her whole life. The character may feel her face and body have betrayed her and she begins to fear the mirror and hate her own eyes and smile. The plain, lined look of gray discovery only tells her she's old and men no longer "see" her. In marked contrast, looking in the mirror when she was young would have told her she was beautiful, youthful, attractive, worthwhile. It is a big ego boost to have the world look at you and its mouth water. The eyes of strangers once undressed her and men wanted her. Now they pet her dog. The bubble has burst.

The great paradox is that while she was young and pretty she was also described as miserable and poor and she spent her time longing for a husband, house, and children. Now that she has them she has traded in her youth and longs to be noticed again. What has she become? The woman is a lonely, bored wife who once was pretty and finds growing old to be painful but inevitable. She misses her sons and daughter who are away at school and no longer need her. She misses her husband who is always at work. She is left with the maid and the dog to mundane, unvarying days without the rosy glow of a bubble to encourage her. She fears death and fears her life will change as she changes.

The funeral of a friend is so significant because the

speakers fears are all capsulized. She sees in her friend's body the betrayal she feels by her own body. Her friend's cold, made-up face is granite among its flowers. Her body has grown old and failed. Inside the friend felt young and thought of herself as exceptional--exclaiming about all she had. Yet, the invincibility of youth is gone and it does not matter if you have everything, no one really has anything. No one is exceptional and everyone more or less trades joy and cheer for all--including a commonplace and sometimes solitary existance ending in a grave.

Jarrell's poetry captures well the phenomenon of aging for a beautiful but ordinary woman. Although I am only twenty-one myself I have seen my mother trying to deal with wrinkles and weight gain and a general middle-age deterioration of the body. People may tell her she is a good cook or a wonderful teacher or mother, but men no longer whistle at her when she walks down the street or tell her she's beautiful.

I realize the phenomenon of a trade off of joy for all in my own life. I married at nineteen thinking that a husband and home and eventually children were what I wanted. Yet, bubbles do burst in real life as well as in poetry and I was left with a bit of soap and water on my hands and a pile of dirty dishes in the sink. We're still together and have been married over two years, but the loss of flirtation bothered me to the point where I stopped wearing my wedding ring.

Ironically, I still have youth as well as a house and husband and I'm trying more or less successfully to keep both. Men still whistle at me and undress me with their eyes and three times I've had lovers. Yet someday I too will grow old and people will stop looking at me. And, I can imagine I will feel then much like the speaker does. My hope of escape from the coming commonplace and solitary years is the memory of today.

When the chief reader asked for scores, she was surprised by the results. Thirty readers scored it a "6," but thirty others gave it a "5," and four more saw it as only a "4." The chief reader was faced with a crucial decision: how to handle a situation in which over half the readers give a "4" or a "5" to a paper judged by the leaders as a strong "6"? The chief reader's decision will critically affect the day's reading: if she allows the scoring to stand as a 5/6--

midway between a "5" and a "6"--the scoring criteria will no longer hold, and rare "6" papers will become virtually extinct. The chief reader knows, in addition, that on the first pair of samples some disagreement is understandable and acceptable. But as the training session moves into its second and third pairs, she must begin "pushing" readers toward consensus.

In what becomes a genuine rhetorical challenge, the chief reader must gently but firmly persuade dissenting readers that the paper is indeed a "6." In this instance, the chief reader had the full support of the committee members. They had been unanimous not only in evaluating the paper as a "6" but in awarding it a high "6." The six leaders further agreed that it was a "top of the range" paper--one representing the best writing Miami's first-year students can do in a two-hour testing situation.

The chief reader opened discussion by asking for comments from those who had scored the paper a "6." She hoped that readers would praise the paper today in much the same way the leaders had yesterday. As if scripted, the readers immediately defended the paper as a "6." In fact, one of the department's most experienced and respected teachers said he thought the paper would be among the very best he would read all day. Heads nodded in agreement. As other readers described the paper as "striking," "insightful," and "powerful," the chief reader began to sense that the "5" readers were being swayed. When she finally asked them to

comment, their responses sounded unconvincing; one reader said only that she thought the paper "stretched the soap thing a little too far." The chief reader ended discussion by announcing that the leaders--those who had read over 200 papers among them--were unanimous in scoring the paper a strong "6."

Because readers had thus far achieved reasonable agreement or shown themselves willing to be persuaded by the judgments of their respected colleagues, the chief reader felt reasonably comfortable as the group turned to a third pair of sample papers. But a feeling of security sometimes spells disaster in a training session. Unexpected responses may come without warning because, as the following paper demonstrates, the scoring done by leaders is not foolproof.

SAMPLE PAPER "I"

In the poem "Next Day" by Randall Jarrell, the first person narration brings forth the character's emotions to create a clear impression of the speaker. The author uses voice, figurative language, and sentence structure to develop this impression. Randall Jarrell creates a stereotypical middle-aged woman who is confused and unhappy with her life.

The speaker is involved in the routine task of grocery shopping. Right at the beginning of the poem the reader can see that the speaker is a stereotype. This is evident because of the word choice in the fourth and fifth lines of the poem. The descriptive phrase "slacked or shorted, basketed, identical food gathering flocks" paints a picture of identical robot like figures who unenthusiastically engage in a routine task. The word "selves" in the sixth line of the poem implies that the speaker considers the other shoppers to be exactly like her. In the tenth line the author points our that the speaker has a station wagon. This adds to the stereotypical image of a middle aged housewife. In the twenty-seventh line of the poem the author describes the grocery boy patting the speaker's dog. The fact that the woman has a dog also adds to the housewife stereotype.

Throughout the poem it is obvious that the speaker is

unhappy with her dull, routine life. She wants the grocery boy to notice her, and it bewilders her that he does not. She wishes that she were young and pretty again. However, when the speaker was young she was also unhappy. The phrase "when I was young and miserable and pretty and poor" makes this evident. When she was young she wished for a husband, a house, and children. Now that she has had all these things, she wishes for what she used to have: her youth and her beauty. This makes the speaker's confusion evident.

Later in the poem the speaker shows that she is afraid. She complains of the "sure unvarying days" of her life; however, she says that she is afraid that her life will change. This seems contradictory, but the speaker is afraid that her life will change for the worse. Then she would be left longing for these days.

The structure of the poem adds to the tone. The poem is written in a strict, repetitious pattern. This pattern is like the dull routine pattern of the speaker's life. It does not flow freely, but it trudges along. Each stanza is patterned like all of the rest, as each day of the speaker's life is patterned as all of the rest.

Using word choice, voice, and sentence structure the author developes a clear impression of the speaker. By studying these aspects and interpreting the poem, the reader gains a clear understanding of the speaker and her feelings.

When readers reported their evaluation of this paper, the chief reader was faced with a challenge significantly more difficult than with the earlier "6" paper. The challenge arose from the glaring discrepancy between the scores of the leaders and the scores of the readers. The leaders had selected the paper to illustrate a "5," but the readers were almost evenly divided between "4" and "3." In fact, the only readers to raise their hands when the chief reader asked who had scored the paper a "5" were the leaders--and they had been given a "crib sheet" by the chief reader so as not to "forget" the score they had agreed upon during the previous day's sample selection.

What had gone wrong? We're still not sure, but we know that the leaders who had earlier evaluated the paper as a "5" had a change of heart when they reread it during the training session. They now saw the paper as a "4"--and not even a strong "4" at that. The ensuing discussion made clear why the paper fell in the middle rather than the upper range. Although readers praised the writer's attempt to connect style and theme at the end of the essay, they considered the paper "mechanical" and "predictable." They also called attention to the essay's repetitiousness, its weak style (several vague "this" constructions), and its uninspired conclusion.

For the chief reader, the significant difference between scores assigned by the leaders and by the readers presented a delicate rhetorical situation. One of her options was to ignore the leaders' score of "5"; after all, readers were not aware of the discrepancy--only the leaders were. Another option was to announce the difference immediately and try to defend the paper as a "5." Instead, since the chief reader had carefully reread the paper herself and experienced the same doubt as the other leaders, she allowed discussion to center on the paper as a "3" or "4" and only <u>afterwards</u> admitted that the leaders had scored it a "5." When a loud buzz arose among the sixty readers, the chief reader laughed and modestly acknowledged, "But we blew it! You guys are good!" Like precocious students, the readers had surpassed their instructors and proved themselves a well-calibrated

group. There was no talk of a walkout.

Although the chief reader's tact and savvy saved the day, a key question remains: Why did the leaders choose an obviously middle-range paper to represent the upper-range score of "5"? A tape-recording of the selection session suggests two answers. First, the paper had been selected late in the session at a time when "5" papers were being specifically sought. "Here," one of the leaders said, "see if you don't think this is a good '5' paper." The paper was read and, sure enough, it was declared a "5." Second, the paper was chosen when the leaders were looking for a more formal and traditional upper-range paper to balance the humor and light touch of the "Donna Reed" paper. Unfortunately, it turned out to be more of a contrast than leaders had anticipated.

We were reminded, by this embarassing experience, of the central importance of the chief reader in making decisions involving the selection and use of sample papers.

REFERENCES

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18, 65-81.

Coffman, W.E., & Kurfman, D.A. (1968). A comparison of two methods of reading essay examinations. American Educational Research Journal, 5, 99-107.

Cooper, P.L. (1984). The assessment of writing ability: A review of research. Princeton: Educational Testing Service.

Diederich, P.B. (1974). Measuring growth in English. Urbana: National Council of Teachers of English.

Hughes, D.C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, 21, 277-281.

Myers, M. (1980). A procedure for writing assessment and holistic scoring. Urbana: National Council of Teachers of English.

Nold, E.W., & Freedman, S.W. (1977). An analysis of readers' responses to essays. Research in the Teaching of English, 11, 164-174.

Northwest Regional Educational Laboratory. (1980). Writing assessment for the 1980s. Proceedings of a national conference on the assessment of writing. Princeton: Educational Testing Service.

Odell, L., & Cooper, C. (1980). Procedures for evaluating

writing: Assumptions and needed research. College English, 42, 35-44.

O'Donnell, H. (1984). Large scale writing assessment: ERIC digest. Princeton: Educational Testing Service.

Spandel, V., & Stiggins, R.J. (1980). Direct measures of writing skill: Issues and applications. Portland: Northwest Regional Educational Laboratory.

Sweedler-Brown, C.O. (1985). The influence of training and experience on holistic essay evaluations. English Journal, 74, 49-55.

White, E.M. (1984). Holisticism. College Composition and Communication, 35, 400-409.

White, E.M. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.